W. James MacLean (Ed.)

Spatial Coherence for Visual Motion Analysis

LNCS 3667

First International Workshop, SCVMA 2004
Prague, Czech Republic, May 2004
Revised Papers

Springer

W. James MacLean (Ed.)

Spatial Coherence for Visual Motion Analysis

First International Workshop, SCVMA 2004
Prague, Czech Republic, May 15, 2004
Revised Papers

Springer

Volume Editor

W. James MacLean
University of Toronto
Department of Electrical and Computer Engineering
10 King's College Road, Toronto, Ontario, M5S 3G4 , Canada
E-mail: maclean@eecg.toronto.edu

Library of Congress Control Number: 2006922617

CR Subject Classification (1998): I.2.10, I.4.8, I.5, I.3.5, F.2.2

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

ISSN 0302-9743
ISBN-10 3-540-32533-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-32533-8 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11676959 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3667

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

Motion analysis is a central problem in computer vision, and the past two decades have seen important advances in this field. However, visual motion is still often considered on a pixel-by-pixel basis, even though this ignores the fact that image regions corresponding to a single object usually undergo motion that is highly correlated. Further, it is often of interest to accurately measure the boundaries of moving regions. In the case of articulated motion, especially human motion, discovering motion boundaries is non-trivial but an important task nonetheless. Another related problem is identifying and grouping multiple disconnected regions moving with similar motions, such as a flock of geese. Early approaches focused on measuring motion of either the boundaries or the interior, but seldom both in unison. For several years now, attempts have been made to include spatial coherence terms into algorithms for 2- and 3-D motion recovery, as well as motion boundary estimation.

This volume is a record of papers presented at the First International Workshop on Spatial Coherence for Visual Motion Analysis, held May 15th, 2004 in Prague, in conjunction with the European Conference on Computer Vision (LNCS 3021–4). The workshop examined techniques for integrating spatial coherence constraints during motion analysis of image sequences. The papers were revised after the workshop to allow for incorporation of feedback from the workshop.

I would like to thank the program committee for their time and effort in reviewing the submissions received for the workshop. Further thanks go to Radim Sara of the ECCV 2004 organizing committee for handling the local arrangements for the workshop. Finally, I would also like to gratefully acknowledge the financial support of MD Robotics, Brampton, Canada.

W. James MacLean

Program Committee

W. James MacLean, University of Toronto (Program Chair)
P. Anandan, Microsoft Research
Andrew Blake, Microsoft Research
Patrick Bouthemy, IRISA/INRIA Rennes
Brendan Frey, University of Toronto
David Fleet, University of Toronto
Allan Jepson, University of Toronto
Takeo Kanade, Carnegie Mellon University
Hans-Hellmut Nagel, Universität Karlsruhe (TH)
Harpreet S. Sawhney, Sarnoff Corporation
Nikos Paragios, Siemens Corporate Research
Hai Tao, University of California, Santa Cruz
Yair Weiss, The Hebrew University of Jerusalem

Lecture Notes in Computer Science

For information about Vols. 1–3821

please contact your bookseller or Springer

Vol. 3927: J. Hespanha, A. Tiwari (Eds.), Hybrid Systems: Computation and Control. XII, 584 pages. 2006.

Vol. 3925: A. Valmari (Ed.), Model Checking Software. X, 307 pages. 2006.

Vol. 3924: P. Sestoft (Ed.), Programming Languages and Systems. XII, 343 pages. 2006.

Vol. 3923: A. Mycroft, A. Zeller (Eds.), Compiler Construction. XV, 277 pages. 2006.

Vol. 3922: L. Baresi, R. Heckel (Eds.), Fundamental Approaches to Software Engineering. XIII, 427 pages. 2006.

Vol. 3921: L. Aceto, A. Ingólfsdóttir (Eds.), Foundations of Software Science and Computation Structures. XV, 447 pages. 2006.

Vol. 3920: H. Hermanns, J. Palsberg (Eds.), Tools and Algorithms for the Construction and Analysis of Systems. XIV, 506 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006. (Sublibrary LNBI).

Vol. 3915: R. Nayak, M.J. Zaki (Eds.), Knowledge Discovery from XML Documents. VIII, 105 pages. 2006.

Vol. 3907: F. Rothlauf, J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J.H. Moore, J. Romero, G.D. Smith, G. Squillero, H. Takagi (Eds.), Applications of Evolutionary Computing. XXIV, 813 pages. 2006.

Vol. 3906: J. Gottlieb, G.R. Raidl (Eds.), Evolutionary Computation in Combinatorial Optimization. XI, 293 pages. 2006.

Vol. 3905: P. Collet, M. Tomassini, M. Ebner, S. Gustafson, A. Ekárt (Eds.), Genetic Programming. XI, 361 pages. 2006.

Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III. XII, 245 pages. 2006. (Sublibrary LNAI).

Vol. 3903: K. Chen, R. Deng, X. Lai, J. Zhou (Eds.), Information Security Practice and Experience. XIV, 392 pages. 2006.

Vol. 3901: P.M. Hill (Ed.), Logic Based Program Synthesis and Transformation. X, 179 pages. 2006.

Vol. 3899: S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. XIV, 216 pages. 2006. (Sublibrary LNAI).

Vol. 3897: B. Preneel, S. Tavares (Eds.), Selected Areas in Cryptography. XI, 371 pages. 2006.

Vol. 3896: Y. Ioannidis, M.H. Scholl, J.W. Schmidt, F. Matthes, M. Hatzopoulos, K. Boehm, A. Kemper, T. Grust, C. Boehm (Eds.), Advances in Database Technology - EDBT 2006. XIV, 1208 pages. 2006.

Vol. 3895: O. Goldreich, A.L. Rosenberg, A.L. Selman (Eds.), Theoretical Computer Science. XII, 399 pages. 2006.

Vol. 3894: W. Grass, B. Sick, K. Waldschmidt (Eds.), Architecture of Computing Systems - ARCS 2006. XII, 496 pages. 2006.

Vol. 3890: S.G. Thompson, R. Ghanea-Hercock (Eds.), Defence Applications of Multi-Agent Systems. XII, 141 pages. 2006. (Sublibrary LNAI).

Vol. 3889: J. Rosca, D. Erdogmus, J.C. Príncipe, S. Haykin (Eds.), Independent Component Analysis and Blind Signal Separation. XXI, 980 pages. 2006.

Vol. 3888: D. Draheim, G. Weber (Eds.), Trends in Enterprise Application Architecture. IX, 145 pages. 2006.

Vol. 3887: J.R. Correa, A. Hevia, M. Kiwi (Eds.), LATIN 2006: Theoretical Informatics. XVI, 814 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006. (Sublibrary LNBI).

Vol. 3885: V. Torra, Y. Narukawa, A. Valls, J. Domingo-Ferrer (Eds.), Modeling Decisions for Artificial Intelligence. XII, 374 pages. 2006. (Sublibrary LNAI).

Vol. 3884: B. Durand, W. Thomas (Eds.), STACS 2006. XIV, 714 pages. 2006.

Vol. 3881: S. Gibet, N. Courty, J.-F. Kamp (Eds.), Gesture in Human-Computer Interaction and Simulation. XIII, 344 pages. 2006. (Sublibrary LNAI).

Vol. 3880: A. Rashid, M. Aksit (Eds.), Transactions on Aspect-Oriented Software Development I. IX, 335 pages. 2006.

Vol. 3879: T. Erlebach, G. Persinao (Eds.), Approximation and Online Algorithms. X, 349 pages. 2006.

Vol. 3878: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. XVII, 589 pages. 2006.

Vol. 3877: M. Detyniecki, J.M. Jose, A. Nürnberger, C. J. '. van Rijsbergen (Eds.), Adaptive Multimedia Retrieval: User, Context, and Feedback. XI, 279 pages. 2006.

Vol. 3876: S. Halevi, T. Rabin (Eds.), Theory of Cryptography. XI, 617 pages. 2006.

Vol. 3875: S. Ur, E. Bin, Y. Wolfsthal (Eds.), Haifa Verification Conference. X, 265 pages. 2006.

Vol. 3874: R. Missaoui, J. Schmidt (Eds.), Formal Concept Analysis. X, 309 pages. 2006. (Sublibrary LNAI).

Vol. 3873: L. Maicher, J. Park (Eds.), Charting the Topic Maps Research and Applications Landscape. VIII, 281 pages. 2006. (Sublibrary LNAI).

Vol. 3872: H. Bunke, A. L. Spitz (Eds.), Document Analysis Systems VII. XIII, 630 pages. 2006.

Vol. 3870: S. Spaccapietra, P. Atzeni, W.W. Chu, T. Catarci, K.P. Sycara (Eds.), Journal on Data Semantics V. XIII, 237 pages. 2006.

Vol. 3869: S. Renals, S. Bengio (Eds.), Machine Learning for Multimodal Interaction. XIII, 490 pages. 2006.

Vol. 3868: K. Römer, H. Karl, F. Mattern (Eds.), Wireless Sensor Networks. XI, 342 pages. 2006.

Vol. 3866: T. Dimitrakos, F. Martinelli, P.Y.A. Ryan, S. Schneider (Eds.), Formal Aspects in Security and Trust. X, 259 pages. 2006.

Vol. 3865: W. Shen, K.-M. Chao, Z. Lin, J.-P.A. Barthès (Eds.), Computer Supported Cooperative Work in Design II. XII, 359 pages. 2006.

Vol. 3863: M. Kohlhase (Ed.), Mathematical Knowledge Management. XI, 405 pages. 2006. (Sublibrary LNAI).

Vol. 3862: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 267 pages. 2006. (Sublibrary LNAI).

Vol. 3861: J. Dix, S.J. Hegner (Eds.), Foundations of Information and Knowledge Systems. X, 331 pages. 2006.

Vol. 3860: D. Pointcheval (Ed.), Topics in Cryptology – CT-RSA 2006. XI, 365 pages. 2006.

Vol. 3858: A. Valdes, D. Zamboni (Eds.), Recent Advances in Intrusion Detection. X, 351 pages. 2006.

Vol. 3857: M.P.C. Fossorier, H. Imai, S. Lin, A. Poli (Eds.), Applied Algebra, Algebraic Algorithms and Error-Correcting Codes. XI, 350 pages. 2006.

Vol. 3855: E. A. Emerson, K.S. Namjoshi (Eds.), Verification, Model Checking, and Abstract Interpretation. XI, 443 pages. 2005.

Vol. 3854: I. Stavrakakis, M. Smirnov (Eds.), Autonomic Communication. XIII, 303 pages. 2006.

Vol. 3853: A.J. Ijspeert, T. Masuzawa, S. Kusumoto (Eds.), Biologically Inspired Approaches to Advanced Information Technology. XIV, 388 pages. 2006.

Vol. 3852: P.J. Narayanan, S.K. Nayar, H.-Y. Shum (Eds.), Computer Vision – ACCV 2006, Part II. XXXI, 977 pages. 2006.

Vol. 3851: P.J. Narayanan, S.K. Nayar, H.-Y. Shum (Eds.), Computer Vision – ACCV 2006, Part I. XXXI, 973 pages. 2006.

Vol. 3850: R. Freund, G. Păun, G. Rozenberg, A. Salomaa (Eds.), Membrane Computing. IX, 371 pages. 2006.

Vol. 3849: I. Bloch, A. Petrosino, A.G.B. Tettamanzi (Eds.), Fuzzy Logic and Applications. XIV, 438 pages. 2006. (Sublibrary LNAI).

Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), Constraint-Based Mining and Inductive Databases. X, 401 pages. 2006. (Sublibrary LNAI).

Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyratos, Y. Tanaka (Eds.), Federation over the Web. X, 215 pages. 2006. (Sublibrary LNAI).

Vol. 3846: H. J. van den Herik, Y. Björnsson, N.S. Netanyahu (Eds.), Computers and Games. XIV, 333 pages. 2006.

Vol. 3845: J. Farré, I. Litovsky, S. Schmitz (Eds.), Implementation and Application of Automata. XIII, 360 pages. 2006.

Vol. 3844: J.-M. Bruel (Ed.), Satellite Events at the MoDELS 2005 Conference. XIII, 360 pages. 2006.

Vol. 3843: P. Healy, N.S. Nikolov (Eds.), Graph Drawing. XVII, 536 pages. 2006.

Vol. 3842: H.T. Shen, J. Li, M. Li, J. Ni, W. Wang (Eds.), Advanced Web and Network Technologies, and Applications. XXVII, 1057 pages. 2006.

Vol. 3841: X. Zhou, J. Li, H.T. Shen, M. Kitsuregawa, Y. Zhang (Eds.), Frontiers of WWW Research and Development - APWeb 2006. XXIV, 1223 pages. 2006.

Vol. 3840: M. Li, B. Boehm, L.J. Osterweil (Eds.), Unifying the Software Process Spectrum. XVI, 522 pages. 2006.

Vol. 3839: J.-C. Filliâtre, C. Paulin-Mohring, B. Werner (Eds.), Types for Proofs and Programs. VIII, 275 pages. 2006.

Vol. 3838: A. Middeldorp, V. van Oostrom, F. van Raamsdonk, R. de Vrijer (Eds.), Processes, Terms and Cycles: Steps on the Road to Infinity. XVIII, 639 pages. 2005.

Vol. 3837: K. Cho, P. Jacquet (Eds.), Technologies for Advanced Heterogeneous Networks. IX, 307 pages. 2005.

Vol. 3836: J.-M. Pierson (Ed.), Data Management in Grids. X, 143 pages. 2006.

Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIV, 744 pages. 2005. (Sublibrary LNAI).

Vol. 3834: D.G. Feitelson, E. Frachtenberg, L. Rudolph, U. Schwiegelshohn (Eds.), Job Scheduling Strategies for Parallel Processing. VIII, 283 pages. 2005.

Vol. 3833: K.-J. Li, C. Vangenot (Eds.), Web and Wireless Geographical Information Systems. XI, 309 pages. 2005.

Vol. 3832: D. Zhang, A.K. Jain (Eds.), Advances in Biometrics. XX, 796 pages. 2005.

Vol. 3831: J. Wiedermann, G. Tel, J. Pokorný, M. Bieliková, J. Štuller (Eds.), SOFSEM 2006: Theory and Practice of Computer Science. XV, 576 pages. 2006.

Vol. 3830: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems II. VIII, 291 pages. 2006. (Sublibrary LNAI).

Vol. 3829: P. Pettersson, W. Yi (Eds.), Formal Modeling and Analysis of Timed Systems. IX, 305 pages. 2005.

Vol. 3828: X. Deng, Y. Ye (Eds.), Internet and Network Economics. XVII, 1106 pages. 2005.

Vol. 3827: X. Deng, D.-Z. Du (Eds.), Algorithms and Computation. XX, 1190 pages. 2005.

Vol. 3826: B. Benatallah, F. Casati, P. Traverso (Eds.), Service-Oriented Computing - ICSOC 2005. XVIII, 597 pages. 2005.

Vol. 3824: L.T. Yang, M. Amamiya, Z. Liu, M. Guo, F.J. Rammig (Eds.), Embedded and Ubiquitous Computing – EUC 2005. XXIII, 1204 pages. 2005.

Vol. 3823: T. Enokido, L. Yan, B. Xiao, D. Kim, Y. Dai, L.T. Yang (Eds.), Embedded and Ubiquitous Computing – EUC 2005 Workshops. XXXII, 1317 pages. 2005.

Vol. 3822: D. Feng, D. Lin, M. Yung (Eds.), Information Security and Cryptology. XII, 420 pages. 2005.

Table of Contents

2D Motion Description and Contextual Motion Analysis: Issues and New Models

P. Bouthemy

IRISA / INRIA,
Campus universitaire de Beaulieu,
35042 Rennes cedex, France

Abstract. In this paper, several important issues related to visual motion analysis are addressed with a focus on the type of motion information to be estimated and the way contextual information is expressed and exploited. Assumptions (i.e., data models) must be formulated to relate the observed image intensities with motion, and other constraints (i.e., motion models) must be added to solve problems like motion segmentation, optical flow computation, or motion recognition. The motion models are supposed to capture known, expected or learned properties of the motion field, and this implies to somehow introduce spatial coherence or more generally contextual information. The latter can be formalized in a probabilistic way with local conditional densities as in Markov models. It can also rely on predefined spatial supports (e.g., blocks or pre-segmented regions). The classic mathematical expressions associated with the visual motion information are of two types. Some are continuous variables to represent velocity vectors or parametric motion models. The other are discrete variables or symbolic labels to code motion detection output (binary labels) or motion segmentation output (numbers of the motion regions or layers). We introduce new models, called mixed-state auto-models, whose variables belong to a domain formed by the union of discrete and continuous values, and which include local spatial contextual information. We describe how such models can be specified and exploited in the motion recognition problem. Finally, we present a new way of investigating the motion detection problem with spatial coherence being associated to a perceptual grouping principle.

1 Introduction

Motion is seamlessly perceived by human beings when directly observing a day-life scene, but also when watching films, videos or TV programs, or even various domain-specific image sequences such as meteorological or heart ultrasound ones. However, motion information is hidden in the image sequences supplied by image sensors. It has to be recovered from the observations formed by the image intensities in the successive frames of the sequence.

Assumptions (i.e., *data models*) must be formulated to relate the observed image intensities with motion. When dealing with video, the commonly used data model is the brightness constancy constraint which states that the intensity does not change along the trajectory of the moving point in the image plane (at least, to a short time extent). The motion constraint equation can then be expressed in a differential form that relates

W.J. MacLean (Ed.): SCVMA 2004, LNCS 3667, pp. 1–15, 2006.

the 2D velocity vector, the spatial image gradient and the temporal intensity derivative at any point p in the image. Nevertheless, this enables to locally retrieve one component of the velocity vector only, the so-called normal flow, which corresponds to the aperture problem. Then, other constraints (i.e., *motion models*) must be added. They are supposed to formalize known, expected or learned properties of the motion field, and this implies to somehow introduce spatial coherence or more generally contextual information.

In this paper, several important issues related to visual motion analysis are addressed with a focus on the type of motion information to be estimated and the way contextual information is formulated and exploited. Visual motion information can involve different kinds of mathematical variables. First, we can deal with *continuous variables* to represent the motion field : velocity vectors $\mathbf{w}(p)$ with $\mathbf{w}(p) \in \mathbb{R}^2$, or parametric motion models with parameters $\theta \in \mathbb{R}^d$ with d denoting the number of parameters. Let us note that the latter can be equivalently represented by the model flow vectors $\{\mathbf{w}_\theta(p)\}$ with $\mathbf{w}_\theta(p) \in \mathbb{R}^2$. Second, we can consider *discrete values or symbolic labels* to code motion detection output: binary values $\{0, 1\}$, or motion segmentation output: number n of the motion region or layer with $n \in \{1, ..., N\}$. Furthermore, we will introduce new models, called *mixed-state auto-models*, whose variables belong to a domain formed by the union of discrete and continuous values, and which include local spatial contextual information too. We will describe how such models can be specified and exploited in the motion recognition problem.

Spatial coherence can be formalized by conditional densities defined on local neighborhoods as in Markov Random Field (MRF) models, or equivalently by potentials on cliques as in Gibbs distributions. Another way is to first segment each image into spatial regions according to a given criterion (grey level, colour, texture) and to analyse the motion information over these regions. Perceptual grouping schemes can also be envisaged.

The remainder of the paper is organized as follows. In Section 2, the motion measurements that can be locally computed are briefly recalled and the subsequent needs for complementary constraints or motion models are outlined. Section 3 reviews briefly several MRF-based approaches we developed in the past to deal with the motion segmentation issue stated as a contextual labeling problem involving discrete variables. Section 4 is concerned with the main aspects of optical flow computation using MRF models or more generally relying on energy minimization methods. In that case, continuous motion variables are considered. Motion recognition or classification, and more specifically event detection in video, is addressed in Section 5, requiring the introduction of new contextual models with mixed states. Section 6 describes a new way to address motion detection based on a perceptual grouping principle.

2 Local Motion Measurements

The brightness constancy assumption along the trajectory of a moving point $p(t)$ in the image plane, with $p(t) = (x(t), y(t))$, can be expressed as $dI(x(t), y(t), t)/dt = 0$, with I denoting the image intensity function. By applying the chain rule, we get the well-known motion constraint equation [22, 32]:

$$r(p,t) = \mathbf{w}(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{1}$$

where ∇I denotes the spatial gradient of the intensity, with $\nabla I = (I_x, I_y)$, and I_t its partial temporal derivative. The above equation can be straightforwardly extended to the case where a parametric motion model is considered, and we can write:

$$r_\theta(p,t) = \mathbf{w}_\theta(p,t).\nabla I(p,t) + I_t(p,t) = 0 \ , \tag{2}$$

where θ denotes the vector of motion model parameters. It can be easily derived from equation (1) that the motion information which can be locally recovered at a pixel p is contained in the *normal flow* given by:

$$\nu(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|} \ . \tag{3}$$

It can also be written in a vectorial form: $\boldsymbol{\nu}(p,t) = \frac{-I_t(p,t)}{\|\nabla I(p,t)\|}\boldsymbol{\omega}_{\nabla I}(p,t)$, where $\boldsymbol{\omega}_{\nabla I}$ denotes the unit vector parallel to the intensity spatial gradient. However, it should be clear that the orientation of the normal flow vector does not convey any information on the motion direction, but implicitly on the object texture (for inner points) or on the object shape (for points on the object border). Besides, the normal flow can be computed at the right scale to enforce reliability as explained in [15].

In case of a moving camera and assuming that the dominant image motion is due to the camera motion and can be correctly described by a 2D parametric motion model, we can exhibit the *residual normal flow* given by:

$$\nu_{res}(p,t) = \frac{-DFD_{\hat{\theta}}(p,t)}{\|\nabla I(p,t)\|} \ , \tag{4}$$

where $DFD_{\hat{\theta}}(p,t) = I(p + \mathbf{w}_{\hat{\theta}}, t+1) - I(p,t)$ is the displaced frame difference corresponding to the compensation of the dominant motion described by the estimated motion model parameters $\hat{\theta}$.

Since the computation of intensity derivatives is usually affected by noise and can be unreliable in nearly uniform areas, it may be preferable to consider the local mean of the absolute magnitude of normal residual flows weighted by the square of the norm of the spatial intensity gradient (as proposed in [23, 36]):

$$\bar{\nu}_{res}(p,t) = \frac{\sum_{q\in\mathcal{F}(p)} \|\nabla I(q,t)\|.|DFD_{\hat{\theta}_t}(q)|}{\max\left(\eta^2, \sum_{q\in\mathcal{F}(p)} \|\nabla I(q,t)\|^2\right)} \ , \tag{5}$$

where $\mathcal{F}(p)$ is a local spatial window centered in pixel p (typically a 3×3 window), and η^2 is a predetermined constant related to the noise level. An interesting property of the local motion quantity $\bar{\nu}_{res}(p)$ is that the reliability of the conveyed motion information can be locally evaluated. Given the lowest motion magnitude δ to be detected, we can derive two bounds, $l_\delta(p)$ and $L_\delta(p)$, verifying the following properties [36]. If $\bar{\nu}_{res}(p) < l_\delta(p)$, the magnitude of the (unknown) true velocity vector $\mathbf{w}(p)$ is necessarily lower than δ. Conversely, if $\bar{\nu}_{res}(p) > L_\delta(p)$, $\|\mathbf{w}(p)\|$ is necessarily greater than δ.

The two bounds l_δ and L_δ can be directly computed from the spatial derivatives of the intensity function within the window $\mathcal{F}(p)$.

By defining the motion quantity $\bar{v}_{res}(p)$, we already advocate the interest of considering spatial coherence to compute motion information. Here, it simply amounts to a weighted averaging over a small spatial support and it only concerns the data model. In the same vein, more information can be locally extracted by considering small spatio-temporal supports, either through spatio-temporal (frequency-based) velocity-tuned filters as in [16] or using 3D orientation tensors [4, 33]. On the other hand, more benefit can be gained by introducing contextual information through the motion models.

3 Discrete Motion Labels and Motion Segmentation

One important step ahead in solving the motion segmentation problem was to formulate the motion segmentation problem as a statistical contextual labeling problem or in other words as a discrete Bayesian inference problem [7, 31]. Segmenting the moving objects is then equivalent to assigning the proper (symbolic) label (i.e., the region number) to each pixel in the image. The advantages are mainly two-fold. Determining the support of each region is then implicit and easy to handle: it merely results from extracting the connected components of pixels with the same label. Introducing spatial coherence can be straightforwardly (and locally) expressed by exploiting MRF models.

Here, by motion segmentation, we mean the competitive partitioning of the image into motion-based homogeneous regions. Motion detection can be viewed as a simplified case where two labels only are considered: static background versus moving object, either with a static camera [1, 30, 39], or a mobile one [36]. The latter assumes that the camera motion (or more specifically, the dominant global motion) can be computed and somehow canceled, usually requiring to resort to robust estimation as we proposed in [35] (joint work with Jean-Marc Odobez). This formulation can also encompass the determination of motion layers by assuming that the regions of same label are not necessarily connected [41].

Formally, we have to determine the hidden discrete motion variables (i.e., region numbers) $l(i)$ where i denotes a site (usually, a pixel of the image grid; it could be also an elementary block [7, 13]). Let $l = \{l(i), i \in S\}$. Each label $l(i)$ takes its value in the set $\Lambda = \{1, .., N_{reg}\}$ where N_{reg} is also unknown. Moreover, the motion of each region is represented by a motion model (usually, a 2D affine motion model of parameters θ which have to be conjointly estimated; we have also explored a non-parametric motion modeling in [13], joint work with Ronan Fablet). Let $\Theta = \{\theta_k, k = 1, .., N_{reg}\}$. The data model of relation (2) is used. The *a priori* on the motion label field (i.e., spatial coherence) is expressed by specifying a MRF model (the simplest choice is to favour the configuration of the same two labels on the two-site cliques so as to yield compact regions with regular boundaries). Adopting the Bayesian MAP criterion is then equivalent to minimizing an energy function E whose expression can be written in the general following form:

$$E(l, \Theta, N_{reg}) = \sum_{i \in S} \rho_1[r_{\theta_{l(i)}}(i)] + \sum_{i \sim j} \rho_2[l(i), l(j)] \ , \tag{6}$$

where $i \sim j$ designates a two-site clique. In [7] (joint work with Edouard François), we considered the quadratic function $\rho_1(x) = x^2$ for the data-driven term in (6). The minimization of the energy function E was carried out on l and Θ in an iterative alternate way, and the number of regions N_{reg} was determined by introducing an extraneous label and using an appropriate statistical test. In [37] (joint work with Jean-Marc Odobez), we instead chose a robust estimator for ρ_1. This allowed us to avoid the alternate minimization procedure and to determine or update the number of regions through an outlier process in every region.

Specifying (simple) MRF models at a pixel level (i.e., sites are pixels and a 4- or 8-neighbour system is considered) is efficient, but remains limited to express more sophisticated properties on region geometry (e.g., more global shape information [10]) or to handle extended spatial interaction. Multigrid MRF models [21] (as used in [36, 37]) is a means to address somewhat the second concern (and also to speed up the minimization process while usually supplying better results). An alternative is to first segment the image into spatial regions (based on grey level, colour or texture) and to specify a MRF model on the resulting graph of adjacent regions as we did in [17] (joint work with Marc Gelgon). The motion region labels are then assigned to the nodes of the graph (which are the sites considered in that case). This allowed us to exploit more elaborated and less local *a priori* information on the geometry of the regions and their motion [17]. However, the spatial segmentation stage is often time consuming, and getting an effective improvement on the final motion segmentation accuracy remains questionable. Using the level-set framework is another way to precisely locate region boundaries while dealing with topology changes [38, 39], but handling a competitive motion partioning of the image (with the number of regions *a priori* unknown) remains an open issue in that context even if recent attempts have been reported [11, 26].

Finally, let us mention other recent work on Bayesian motion segmentation, exploring the use of edge motion [42], offering extension to spatio-temporal models [11], or introducing (two-step) hidden Markov measure field (HMMF) models [27]. Tensor voting could also be considered as an implicit way to enforce spatial coherence [34].

4 Continuous Motion Information and Optical Flow Computation

By definition, the velocity field formed by continuous vector variables is a complete representation of the motion information. Computing optical flow based on the data model of equation (1) requires to add a motion model enforcing the expected spatial properties of the motion field, that is, to resort to a regularization method. Such properties of spatial coherence (more specifically, piecewise continuity of the motion field) can be expressed on local spatial neighborhoods. First methods to estimate discontinuous optical flow fields were based on MRF models associated with Bayesian inference [20, 30, 43] (i.e., minimization of a discretized energy function). Then, continuous-domain models were designed based on PDE formalism [2, 8, 25, 46]. Spatial coherence can also be explicitly formulated by first segmenting the image in spatial regions forming the delimited domains where motion models, either dense or parametric ones, can be defined and estimated [6, 17].

A general formulation of the global (discretized) energy function to be minimized to estimate the velocity field $\mathbf{w}$ can be given by:

$$E(\mathbf{w},\zeta) = \sum_{p\in S} \rho_1[r(p)] \; + \sum_{p\sim q} \rho_2[\|\mathbf{w}(p)-\mathbf{w}(q)\|, \zeta(p'_{p\sim q})] + \sum_{A\in\chi} \rho_3(\zeta_A) \;\; , \quad (7)$$

where S designates the set of pixel sites, $r(p)$ is defined in (1), $S' = \{p'\}$ the set of discontinuity sites located midway between the pixel sites and χ is the set of cliques associated with the neighborhood system chosen on S'. In [20] (joint work with Fabrice Heitz), quadratic functions were used and the motion discontinuities were handled by introducing a binary line process ζ. Then, robust estimators were popularized [5, 28] leading to the introduction of so-called auxiliary variables ζ now taking their values in $[0, 1]$. Depending on the followed approach, the third term of the energy $E(\mathbf{w}, \zeta)$ can be optional. Multigrid MRF are moreover involved in the scheme developed by Mémin and Pérez in [28]. Besides, multiresolution incremental schemes are required to compute optical flow in case of large displacements. Dense optical flow and parametric motion models can also be jointly considered and estimated, which enables to supply a segmented velocity field as designed by Mémin and Pérez [29].

Recent advances have dealt with the computation of fluid motion fields involving the definition of a new data model (derived from the continuity equation of the fluid mechanics) and of a motion model preserving the underlying physics of the visualized fluid flows (2^{nd} order div-curl constraint) as defined by Corpetti, Mémin and Pérez in [9]. A comprehensive investigation of physics-based data models is described in [19].

5 Motion Recognition and Mixed-State Auto-models

5.1 Event Detection in Video and Mixed-State Probabilistic Models

A big challenge in computer vision consists in approaching the "semantic" content of video documents while dealing with physical image signals and numerical measurements. Here, we consider the detection of relevant events (dynamic content). Therefore, we focuse on motion information and we propose new probabilistic image motion models. The motion information is captured through low-level motion measurements so that it can be efficiently and reliabily computed in any video whatever its genre and its content. Our approach (joint work with Gwénaëlle Piriou and Jian-Feng Yao [40]) consists in modeling separately the camera motion (i.e., the dominant image motion) and the scene motion (i.e., the residual image motion) in a sequence, since these two sources of motion bring important and complementary information. The dominant image motion is represented by a deterministic 2D affine motion model (which is a usual choice):

$$\mathbf{w}_\theta(p) = (a_1 + a_2 x + a_3 y, a_4 + a_5 x + a_6 y)^T \;\; , \quad (8)$$

where $\theta = (a_i, i = 1, \ldots, 6)$ is the model parameter vector and $p = (x, y)$ is an image point. This simple motion model can handle different camera motions such as panning, zooming, tracking, (including of course static shots). To estimate the motion

parameters θ, we employ the robust real-time multi-resolution algorithm[1] described in [35]. The motion model parameters are directly computed from the spatio-temporal derivatives of the intensity function. Consequently, the model motion vector $\mathbf{w}_{\theta_t}(p)$ is available at any pixel p and time t. The two components of $\mathbf{w}_{\theta_t}(p)$ are finely quantized, and we build the empirical 2D histogram of their distribution over the considered video segment. Finally, this histogram is represented by a mixture of 2D Gaussian distributions denoted γ^{cam}. The number of components of the mixture is determined with the Integrated Completed Likelihood criterion (ICL) and their parameters are estimated using the Expectation-Maximization (EM) algorithm [40].

The residual motion measurements are given by the $\bar{\nu}_{res}(p,t)$'s as defined in (5). The probabilistic model of scene motion is derived from global statistics on these measurements. The 1D histograms of $\bar{\nu}_{res}(p,t)$ which have been computed over different video segments, present usually a prominent peak at zero and a continuous component part. The latter can be modeled either by an exponential distribution or a zero-mean Gaussian distribution, both restricted to $]0,\infty[$ (since by definition $\bar{\nu}_{res}(p,t) \geq 0$). Therefore, we consider a specific mixture model to represent the distribution of the local residual motion measurements within a video segment with density [40]:

$$f(z) = \varrho\delta_0(z) + (1-\varrho)\phi_\kappa(z) \ , \tag{9}$$

where z holds for $\bar{\nu}_{res}(p,t)$, ϱ is the mixture weight, δ_0 denotes the Dirac function at 0, and ϕ_κ designates either the (restricted) Gaussian density function with variance $1/2\kappa$ or the exponential density function with mean $1/\kappa$, both with support $]0,\infty[$. Consequently, the proposed model has explicitly two degrees of freedom: ϱ handles the peak at zero and κ accounts for the continuous component of the distribution. ϱ and κ are estimated using the ML criterion. In order to capture not only the instantaneous motion information but also its temporal evolution over the video segment, the temporal contrasts $\Delta\bar{\nu}_{res}$ of the local residual motion measurements are also considered: $\Delta\bar{\nu}_{res}(p,t) = \bar{\nu}_{res}(p,t+1) - \bar{\nu}_{res}(p,t)$. They are modeled, in a similar manner as in (9), by a mixture model $g(z')$ of a Dirac function at 0 and a zero-mean Gaussian distribution, where z' holds for $\Delta\bar{\nu}_{res}(p,t)$. The mixture weight and the variance of the Gaussian distribution are again evaluated using the ML criterion. The full probabilistic residual motion model is then simply defined as the product of these two models: $h^{res}(z,z') = f(z).g(z')$.

Let us stress the peculiar nature of the probabilistic model introduced in relation (9). The value 0 plays a particular role since it accounts for no motion which is a clear semantic information. We can consider that it corresponds to a symbolic state defined by the discrete value $z = 0$ and that the other state is defined by $z > 0$. Therefore, the variable z takes its value in the set $\{0\} \cup]0,\infty[$. We call such a set a *mixed-state space*.

The event detection proceeds in two steps. The first step permits to eliminate the segments that are not likely to contain the searched relevant events. Typically, if we consider sports videos, we try to first distinguish between "play" and "no play" segments. This step is based on the residual motion only. The second step consists in retrieving several specific events among the candidate segments $\{s_0,\ldots,s_N\}$. Here, the two kinds

[1] The corresponding software called MOTION-2D can be downloaded at http://www.irisa.fr/vista/Motion2D.

of motion information (residual and camera motion) are required since the combination allows us to characterize more finely a specific event. A residual motion model with density h_j^{res} and a camera motion model with density γ_j^{cam} have to be previously estimated from a training set of video samples, for each type j of event to detect. The label l_i of each segment s_i is determined using the ML criterion:

$$l_i = \arg \max_{j=1,\ldots,J} \prod_{(p,t)\in s_i} h_j^{res}(z_{(p,t)}, z'_{(p,t)}) \prod_{(p,t)\in s_i} \gamma_j^{cam}(\mathbf{w}_{\hat{\theta}_t}(p,t)) \ . \qquad (10)$$

More details and results on sports videos can be found in [40].

5.2 Mixed-State Auto-models and Motion Classification

Here, we describe joint work with Jian-Feng Yao and Gwénaëlle Piriou and report preliminary results. The scene motion model (to be learnt from image data) defined above only accounts for global (occurrence) statistics accumulated over both the image plane and time (i.e., over all the frames of the video segment). Obviously, it does not capture how the motion information is spatially (or temporally) organized. In [14, 15] (joint work with Ronan Fablet and Patrick Pérez), we have proposed the design of causal Gibbs models from scale and temporal co-occurrences of quantized motion values $\bar{\nu}$. Here, we will extend the model (9) to take into account spatial interaction between neighbours, and define mixed-state auto-models (to follow the terminology introduced in [3]). We will consider the Gaussian case only, but mixed-state auto-models can be defined as well for any distribution from the exponential distribution family [18].

Let us first rewrite the mixed-state probabilistic model (9) in the following exponential family form:

$$f_\theta(z) = \exp\left[\langle \theta, B(z)\rangle - \psi(\theta)\right] \ , \qquad (11)$$

$$\text{with} \quad \theta = (\theta_1, \theta_2)^T = \left(\log \frac{(1-\varrho)\phi_\kappa(0)}{\varrho}, \kappa\right)^T \ , \quad B(z) = (\delta^*(z), -z^2)^T \ ,$$

where $\delta^*(z) = 1 - \delta_0(z)$. Let us note that we can easily recover the original parameters ϱ and κ from the "natural" ones θ_1 and θ_2.

To build our mixed-sate auto-models for the field $(z_i, i \in S)$, we start by considering, as in [3], the family of conditional densities $\mu_i(z_i|\cdot) := \mu_i(z_i|z_j, j \neq i)$, that is the conditional distribution of z_i at a site i given its outside configuration $(\cdot) = (z_j, j \neq i)$. Because of the mixed-state nature of the observations at hand, namely the residual motion measurements, we require that all these conditional distributions are of type defined in (9), or equivalently (11). Let us note that, for each i, the parameters $\theta_i(\cdot) = (\theta_{i,1}(\cdot), \theta_{i,2}(\cdot))$ of the conditional density $\mu_i(z_i|\cdot)$ (here, we use the representation (11)) depend on the spatial context $(\cdot) := (z_j, j \neq i)$. It can be shown [18] that there are vectors $\alpha_i = (a_i, b_i) \in \mathbb{R}^2$ and 2×2 matrices $\beta_{ij} = \begin{pmatrix} c_{ij} & d_{ij} \\ d^*_{ij} & e_{ij} \end{pmatrix}$, such that:

$$\theta_i(\cdot) = \alpha_i + \sum_{j\neq i} \beta_{ij} B(z_j) \ ,$$